Intelligente Informationsextraktion

Technologien & Anwendungen

Dr. Günter Neumann

Forschungsbereich Sprachtechnologie
DFKI, Saarbrücken

Vortrag anlässlich zur Ernennung zum
DFKI Research Fellow am 3. Mai 1999

Die enorme Menge an elektronisch verfügbaren Texten erfordert neuartige Technologien zur Informationsextraktion (IE)

Das Ziel in der IE-Forschung ist die Entwicklung von Methoden zur aufgabenspezifischen Identifikation, Sammlung und Normalisierung von relevanter Information bei gleichzeitigem „Überlesen“ der irrelevanten Information.

Kernfunktionalität eines IE-Systems:

Eingabe:
1. Spezifikation der relevanten Informationen in Form von Templates (getypte Merkmalsstrukturen), z.B. Firmen-, Produktinformation, Personalwechsel, Veranstaltungen
2. Menge von freien Texten (Zeitschriften, Berichte, Internet)

Ausgabe:
Eine Menge von instanziierten Templates, die mit relevanten Textfragmenten gefüllt sind (eventuell normalisiert)

Beispiel für Informationsextraktion 1

Lübeck (dpa) - Die **Lübecker Possehl-Gruppe**, ein im Produktions-, Handel- und Dienstleistungsbereich tätiger Mischkonzern, hat **1994** den **Umsatz** kräftig um **17 Prozent** auf rund **2,8 Milliarden DM gesteigert**. In das neue Geschäftsjahr sei man ebenfalls „mit Schwung“ gestartet. Im **1. Halbjahr 1995** hätten sich die **Umsätze** *des Konzerns* im Vergleich zur Vorjahresperiode um **fast 23 Prozent** auf rund **1,3 Milliarden erhöht**.

Type	C-name	year	amount	tendency	difference
turnover	Possehl1	1994	2.8e+9DM	+	17%
turnover	Possehl1	1995/1	1.3e+9DM	+	23%

Beispiel für Informationsextraktion 2

Ausschnitt aus dem Jahresbericht der RWE (1998):
Eine Schwerpunktregion im Rahmen der Internationalisierung im Energiebereich ist Osteuropa. Hier haben wir unser Engagement im abgelaufenen Geschäftsjahr weiter ausbauen können. Nach dem **Kauf** weiterer **Anteile** halten ***wir*** inzwischen jeweils **knapp über 50%** an den ungarischen Energieversorgungsunternehmen **ELMÜ**, **ÉMÁSZ** und **MÁTRA**. Im Falle von **MÁTRA hat RWE Energie** im **April 1998 Anteile** an **Rheinbraun abgegeben**. Die Präsenz in Polen wurde durch Kooperationsvereinbarungen mit den Regionalversorgern **Zaklad Energetyczny Krakow S.A. (ZEK)** und **Stoleczny Zaklad Energetyczny S.A. (STOEN)** ....im **Frühjahr 1998** weiter ausgebaut.

Group/Subs.	YEAR	KIND	FROM	TO	POT	AMOUNT
RWE	1998	+	ELMÜ			>50%
RWE	1998	+	ÉMÁSZ			>50%
RWE	1998	+	MÁTRA			>50%
RWE Energie	1998	-	MÁTRA	Rheinbraun	4.1998	

Aus Sicht der KI und speziell der Natürlichsprachlichen Verarbeitung (NLP) ist die IE-Forschung sehr attraktiv

- Extraktionsaufgaben sind wohl-definiert
- IE verwendet reale Textdokumente
- IE bedingt schwierige und interessante Probleme für NLP
- IE benötigt systematische Schnittstellen-Spezifikationen zwischen natürlicher Sprache und Domänenwissen
- IE Performanz kann mit menschlicher Performanz verglichen werden (Message Understanding Conferences bzw. TREC)

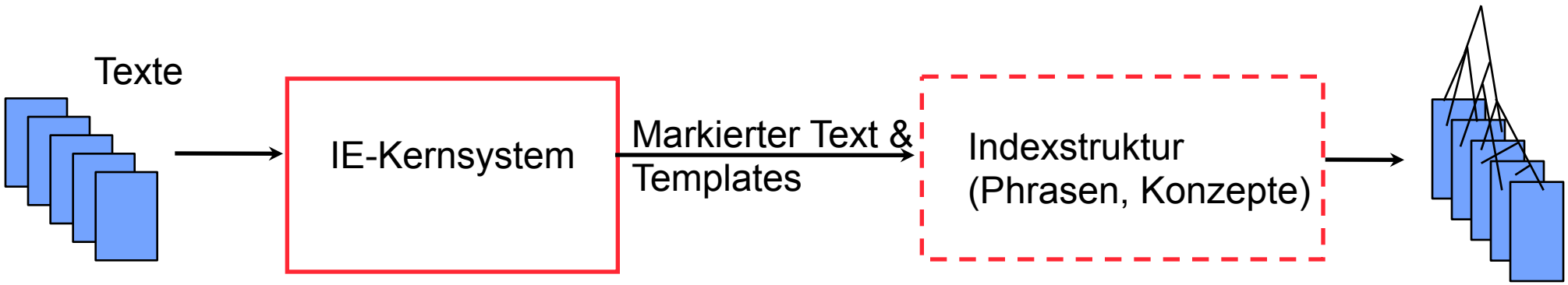
$\Rightarrow$ IE Systeme sind ein Schlüsselfaktor beim Übergang von der Verarbeitung kleiner und künstlicher Daten zur Verarbeitung großer, realer Textmengen (Cowie & Lehnert, 1996)

Hohes Anwendungspotential der Informationsextraktion (IE)

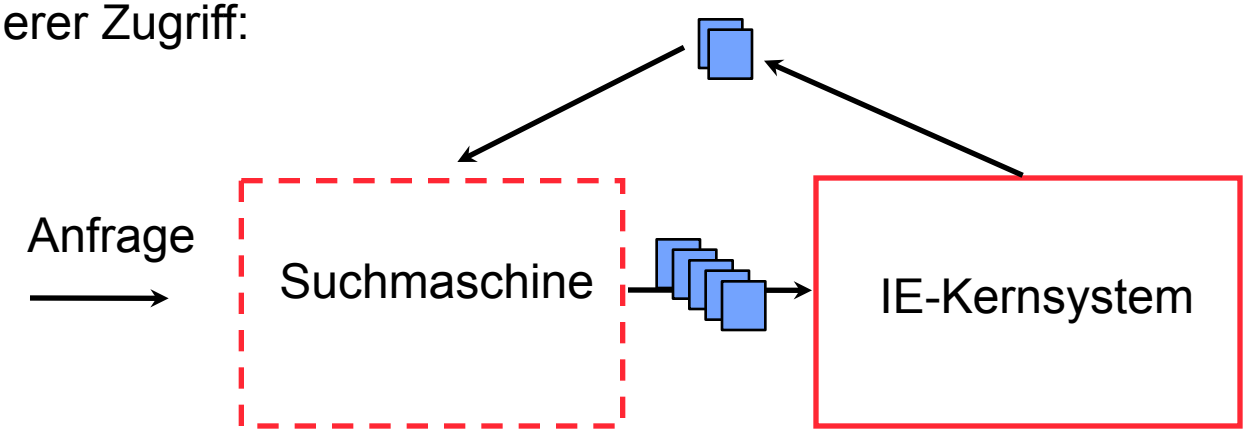
IE und Information-Retrieval	Konstruktion von feinkörnigen Indizes, die stärker an der aktuellen Bedeutung des speziellen Textes orientiert sind
IE und Textklassifikation	Bestimmung von feinkörnigen Entscheidungsregeln
IE und Data Mining	Verbesserte Qualität extrahierter struktureller Information
IE und DBMS	Einsatz in semi-strukturellen DB-Modellen
IE und Wissensbasierte Systeme	Erstellung von Ontologien auf der Basis der extrahierten Information

IE erlaubt eine feinkörnige, kompakte Indizierung (vgl. BMBF Verbundvorhaben Getess)

Kompaktere Indizes:



Feinerer Zugriff:



Beziehung zwischen Informationsextraktion, Data Mining und Text Mining

Informationsextraktion

Extraktion relevanter Information aus unstrukturierten Textdokumenten; Domänenwissen und vorspezifizierte Templates

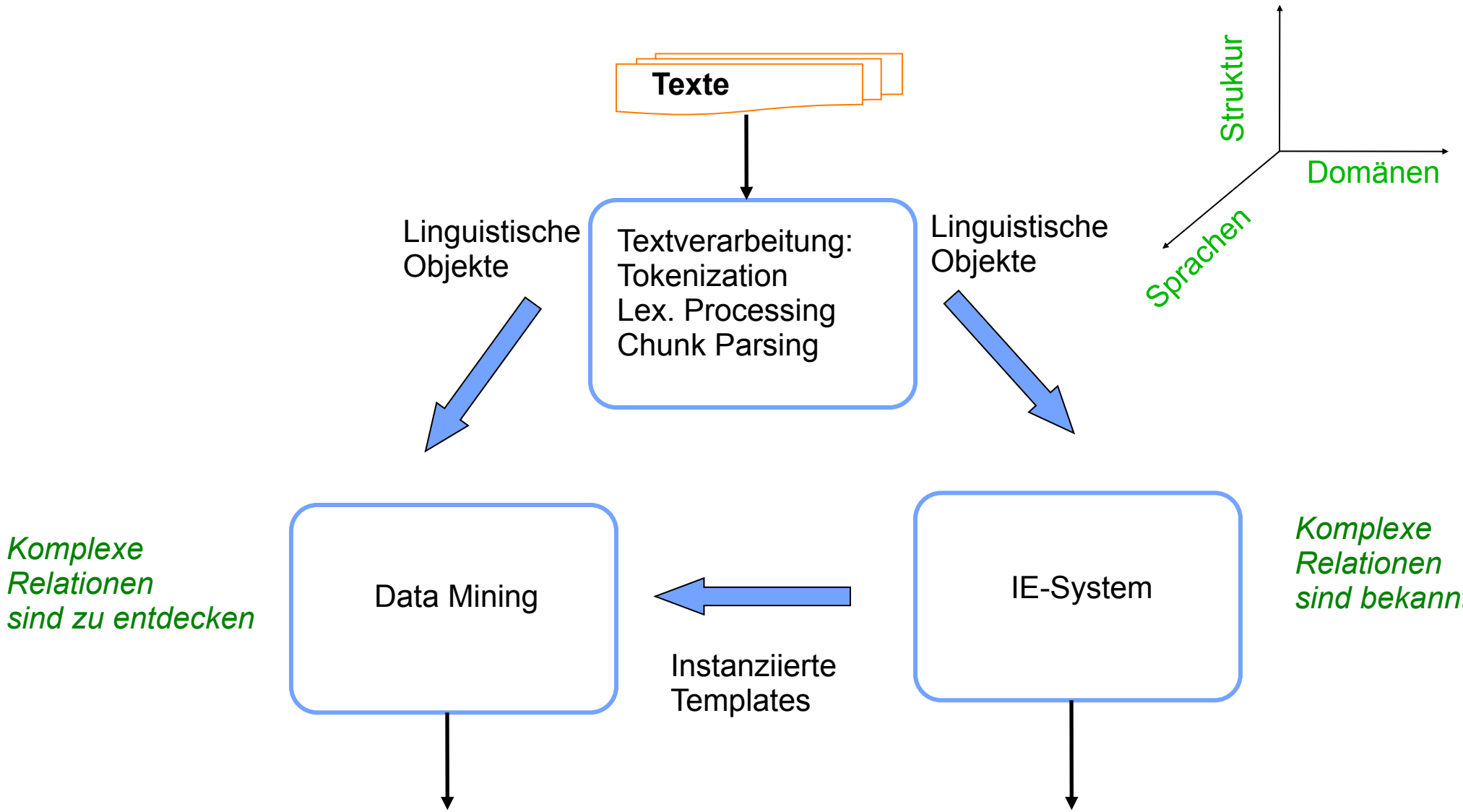
Data Mining

Informationsextraktion aus strukurierten Datenquellen und automatische Entdeckung von relationalen Beziehungen

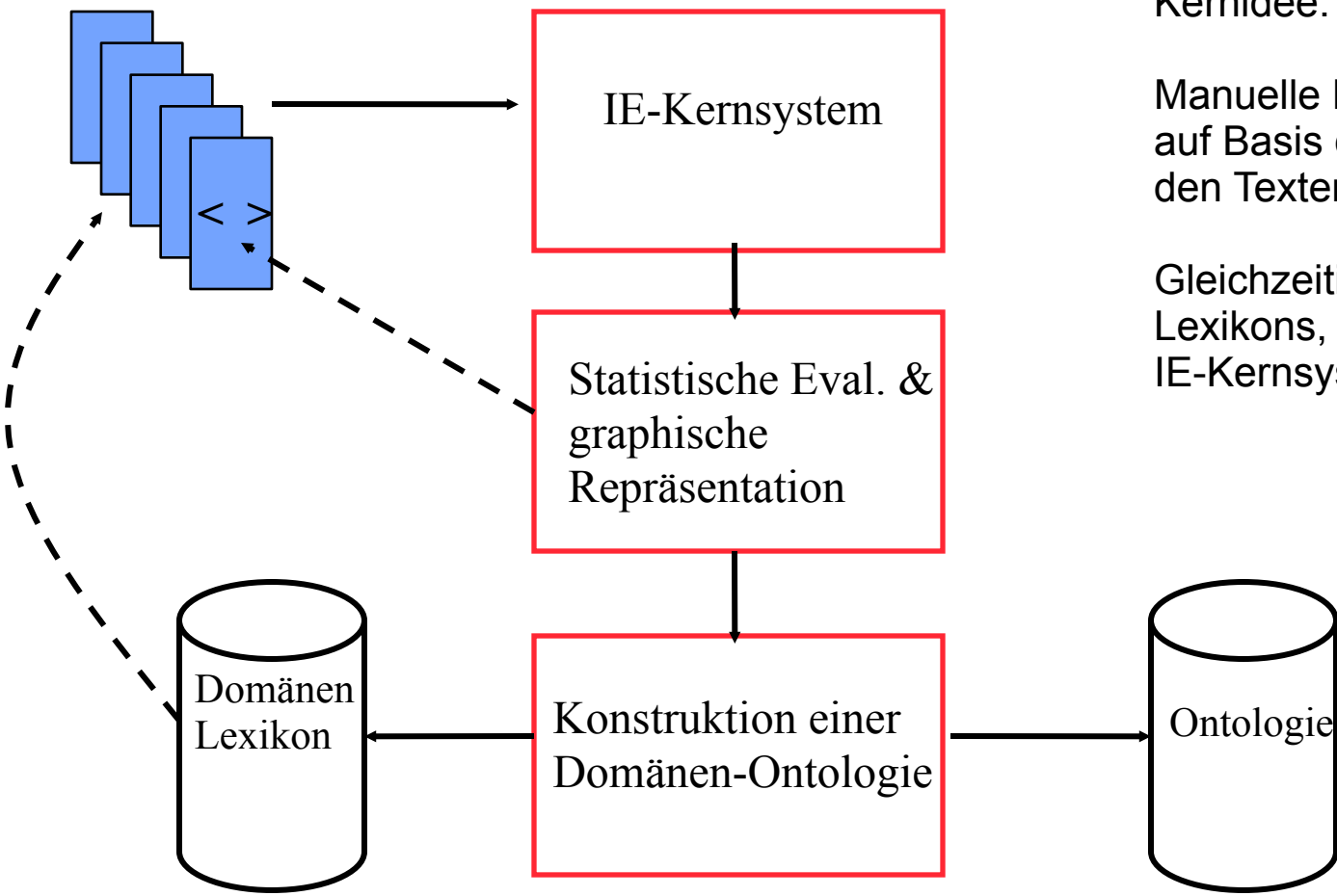
Text Mining

Data mining aus unstrukturierten Textdokumenten unter Verwendung von domänenunabhängiger flacher Textverarbeitung

Flache Textverarbeitung als gemeinsamer Kern in Text Mining und IE



IE als Kernkomponente zur inkrementellen Wissenskonstruktion (vgl. BMBF Verbundvorhaben Getess)



Kernidee:

Manuelle Konstruktion der Domänen-Ontologie auf Basis der linguistischen Information, die aus den Texten extrahiert wurde

Gleichzeitig Konstruktion eines Domänen-Lexikons, welches im nächsten Zyklus im IE-Kernsystem eingesetzt wird (bootstrapping)

Aus Sicht einer Systementwicklung existieren zur Zeit zwei Ansätze

- Sprachtechnologischer Ansatz
  - 😐 linguistisches Wissen manuell durch Experten spezifiziert
  - 😐 Abbildung zwischen NL und Domänenwissen hand-kodiert
  - ☹ manuelle Inspektion des Korpus zur Bestimmung, wie spezifisch Domänenwissen verbalisiert ist
  - ☺ z.Z. der beste Ansatz zur Konstruktion praktikabler Anwendungssysteme
  - ☺ Enwicklung von Werkzeugen zur Unterstützung der Anwendungsentwicklung

- Lern-basierter Ansatz
  - wende möglichst in jedem Teilbereich statistische Methoden an
  - lerne Template-Instanziierungsregeln aus annotierten Daten unter Verwendung von Verfahren des Maschinellen Lernens
  - ☺ zeigen bereits vielversprechende Ergebnisse für einige IE-Teilaufgaben (Eigennamen-Erkennung, flache Instanziierungsregeln)
  - ☺ Abbildung zwischen NL und Domänenwissen wird automatisch induziert
  - ☹ benötigt noch sehr hohen Betrag an annotierten Korpora

Gemeinsam ist beiden Ansätzen der Einsatz von flachen, wiederverwendbaren NL Kernkomponenten (unterschiedlicher Granularität)

- Tokenization, Textscanning/Textzoning (z.B. Analyse von Tabellen, Überschriften, HTML/XML)
  - in den meisten Fällen einfach, aber sehr wichtig

- Morphologische & lexikalische Verarbeitung
  - hoch-abdeckende und sehr schnelle morphologische Analyse
  - Verarbeitung von unbekannten Wörtern und Komposita
  - Part-of-Speech Tagging (Auflösung von mehrdeutigen Wortformen)

- Erkennung von Eigennamen (Personen-, Firmennamen, Datum, Zeit, Maßeinheiten)
  - wichtiger Aspekt hier: kreative Namensbildung, multi-linguale Terme

Einsatz flacher, wiederverwendbarer NL Kernkomponenten zur Analyse von Texten

- Flaches Parsing: sehr lange Sätze (> 30 Wörter), Einbettungen, Aufzählungen
  - Integration von domänenspezifischen Subgrammatiken
  - partielles, kasakadiertes Parsing
  - sehr robuste und effiziente Strategien nötig

- Diskursanalyse: finde unterschiedliche Verbalisierungen desselben Objektes
  - NP-Analyse
  - Koreferenzen
  - relationale Verknüpfungen

Flaches Parsing (Syntaxanalyse)

Tief: [Der [neue [Vertrag [mit den [Klauseln [am Anfang]]]]]] [scheint [am [Beispiel [der [letzten Skizze]]]] [festgelegt worden zu sein]].

⇒ mehrdeutige Klammerung möglich!

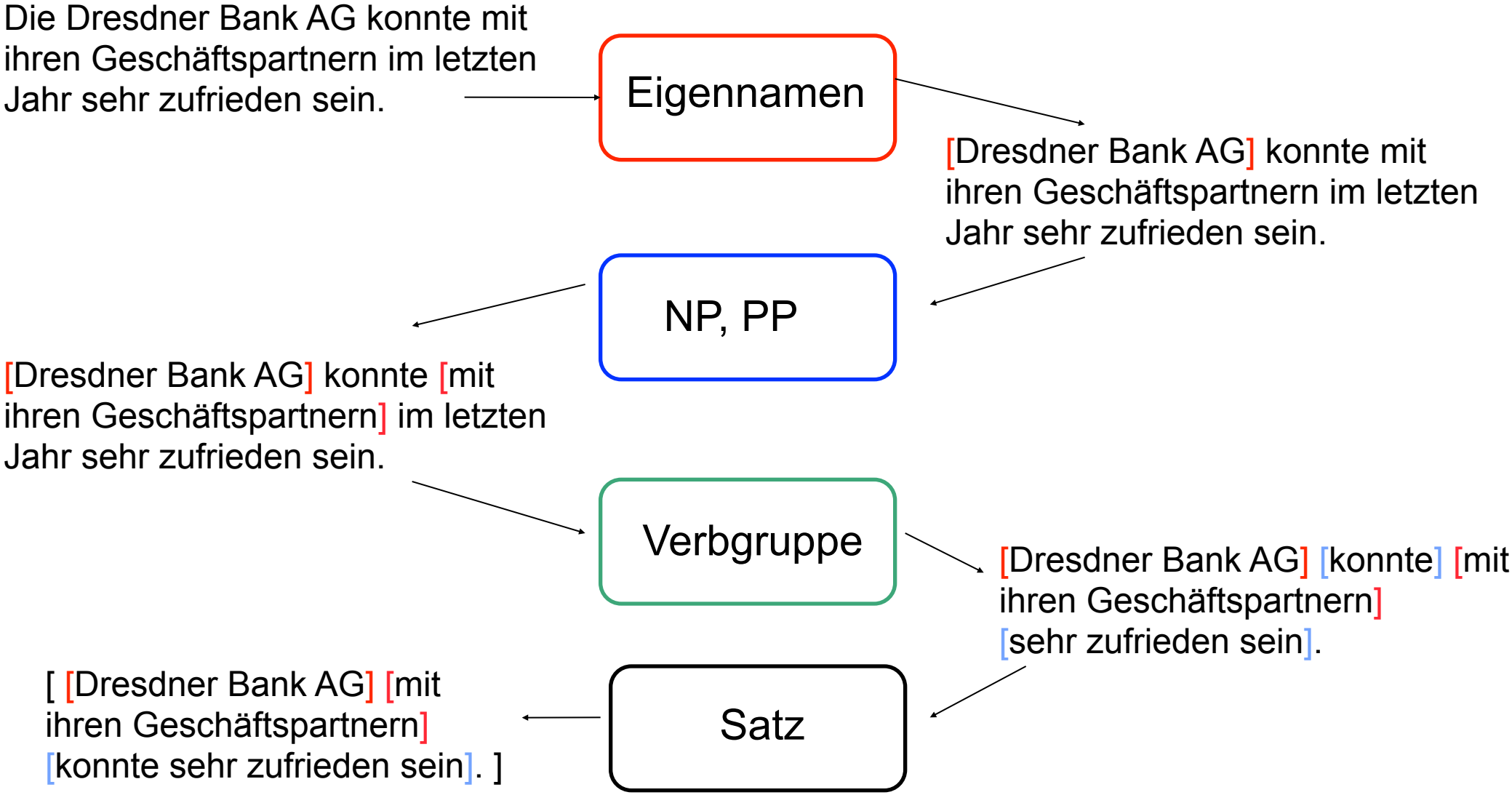
Flach: [[Der neue Vertrag] [mit den Klauseln] [am Anfang]] [scheint] [am Beispiel [der letzten Skizze]] [festgelegt worden zu sein].

⇒ nur grobe Klammerung

Aktuelles Forschungsziel: Integration von flacher und tiefer Analyse

⇒ parametrisierbare Analysetiefe

Partielles Parsing von komplexen Strukturen mittels Kaskaden von endlichen Automaten



Beispiel für Diskursverarbeitung/Koreferenzauflösung

- Beispiele:
  Computec *vermarktet Informationen zu Electronic Games, spieleergänzende Software und Internet-Spiele über eigene Printmedien, CD-ROMs und das Internet. Der Vorstand besteht aus Christian Geltenpoth und Roland Arzberger.* ***Das Unternehmen*** *ist nach eigener Einschätzung der führende Medienanbieter im stark wachsenden Markt der Electronic Games in Deutschland.*

  Da flüchten ***sich die einen*** *ins Ausland, wie etwa* ***der Münchner Strickwaren-hersteller März GmbH*** *oder* ***der badische Strumpffabrikant Arlington Socks,GmbH****. Ab kommendem Jahr strickt* ***März*** *knapp drei Viertel seiner Produktion in Ungarn.*

- Koreferenzauflösung durch Heuristiken wie beispielsweise:

  Die Nominalphrase “das Unternehmen” bezieht sich auf das als letztes genannte Unternehmen im Text.

Metrik zur Beurteilung des Ergebnisses einer IE

- Präzision *P* (Precision): Anteil der korrekt gewonnenen (d.h. der relevanten) Wissensitems im Vergleich zu den insgesamt gefundenen Wissensitems

- Vollständigkeit *V* (Recall) : Anteil der korrekt gewonnenen Wissensitems im Vergleich zu den insgesamt gewinnbaren Wissensitems

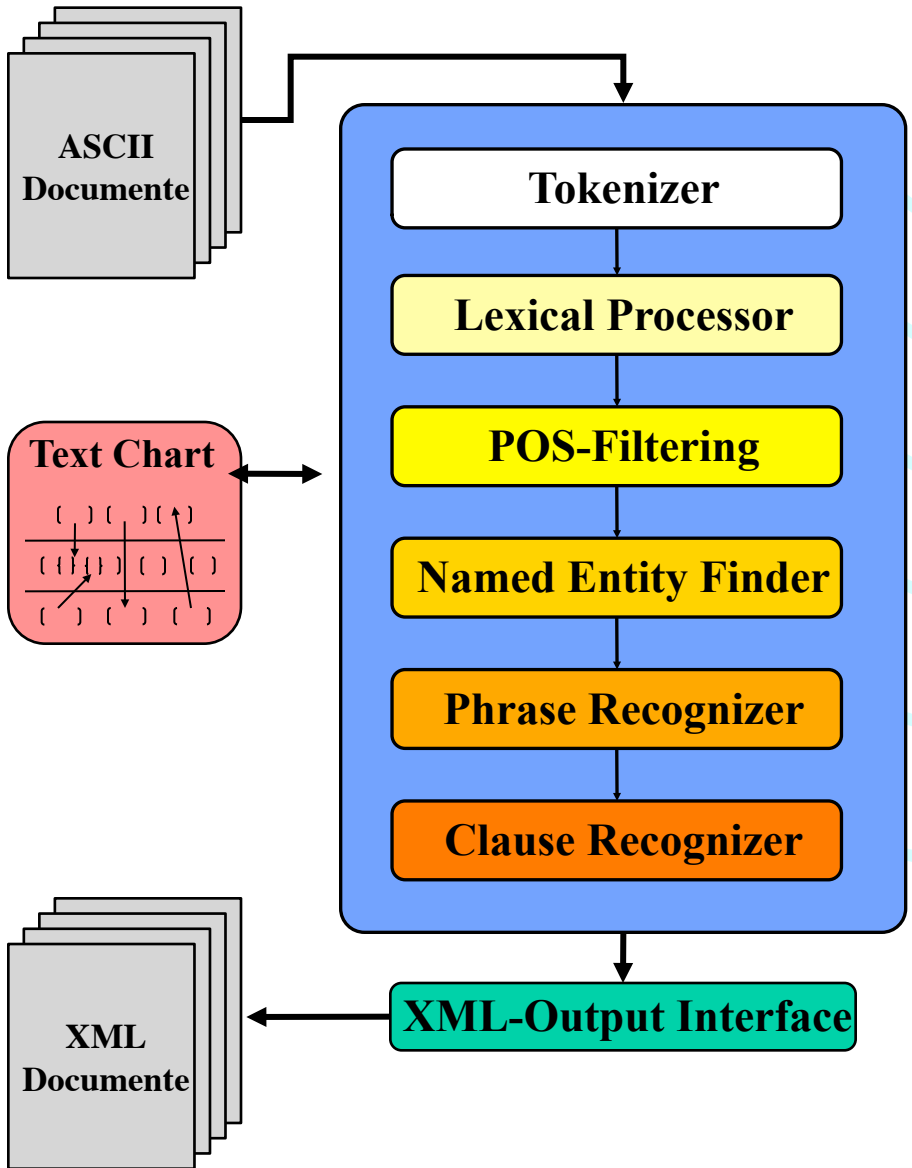
- F-Maß als zusammenfassende Metrik:

$$F = \frac{(\beta^2 + 1) * P * V}{(\beta^2 P + V)}$$

Nur Präzision berücksichtigt: $\beta = 0$
Typischerweise: $\beta = 1$

- **0.6 Barriere** (ab 0.75 Operationalisierung)

Im Rahmen des BMBF-Projektes Paradime wurde am DFKI LT-Lab der sehr leistungsfähige IE-Prototyp SMES/SPPC entwickelt (z.Z. speziell für Deutsch)



Beispiel: rund 60 bis 70 Prozent der Steigerungsrate

rund: low-w
60: 2int

52 Klassen

Steigerungsrate: steigerung+[s]+rate
bis: prep|adv

150.000 Stämme
on-line Komposita
Hyphen-Koordination

bis: adv

> 100 Regeln, Roche&Schabes Ansatz

rund 60 bis 70 Prozent: percentage-NP

12 Subgrammatiken
dynamisches Lexikon
Reference-Resolution

rund 60 bis 70 Prozent: NP
der Steigerungsrate: NP

NPs, PPs, VGs

Divide-and-Conquer Chunk Parser

parametrisierbar

Finite-state Technologie

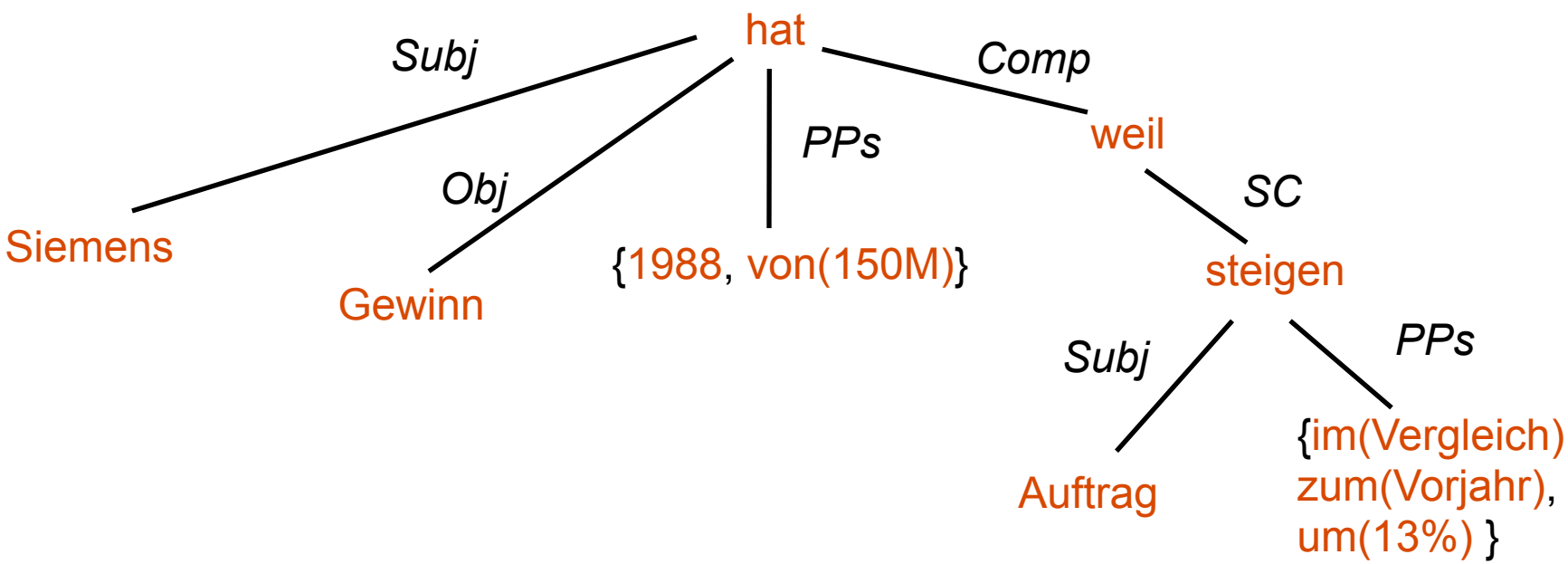
An analysed text is represented as a sequence of underspecified (partial) functional descriptions UFDs

UFD: flat dependency-based structure, only upper bounds for attachment and scoping

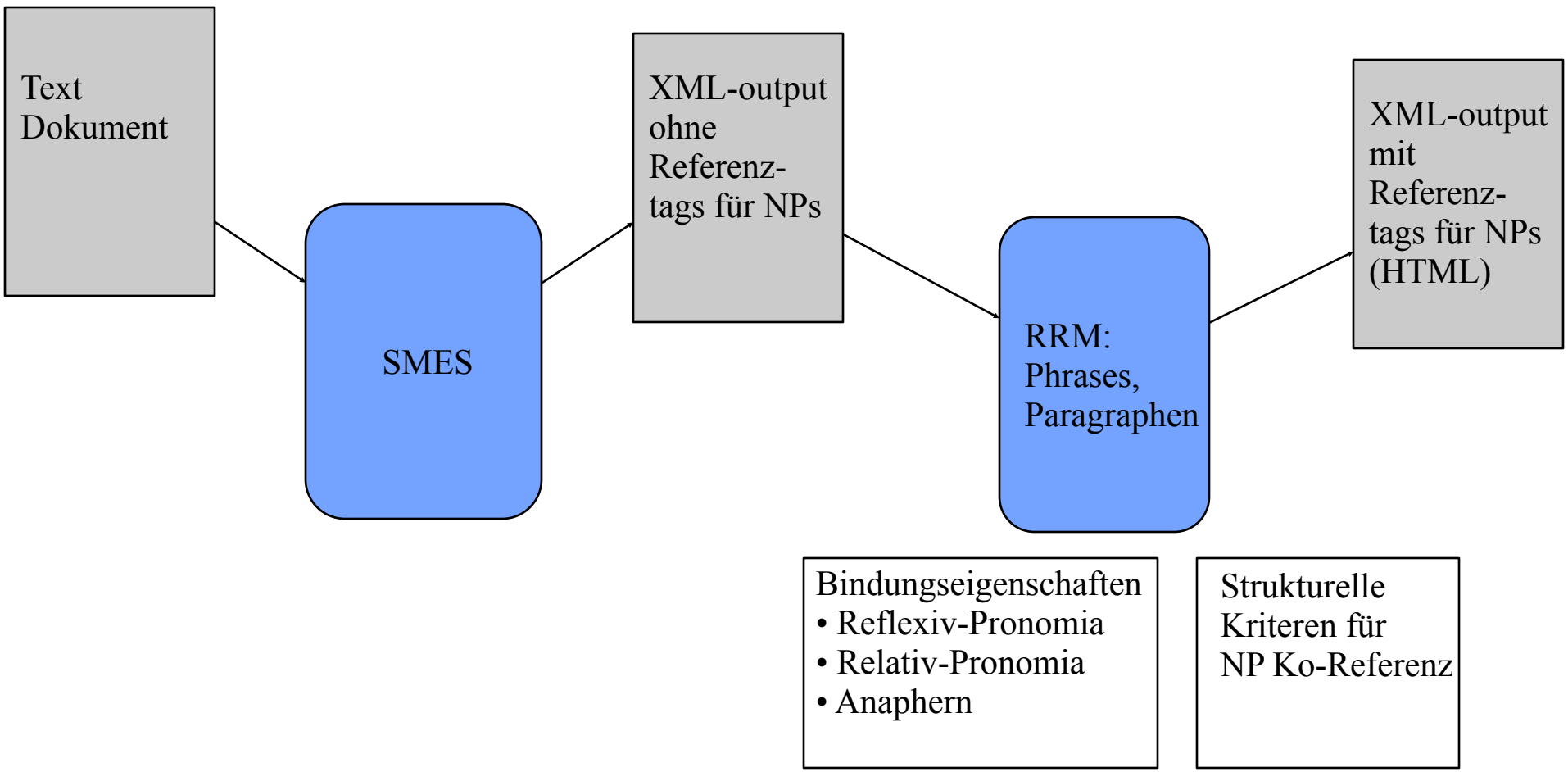
[$_{PN}$Die Siemens GmbH] [$_{V}$hat] [$_{year}$1988][$_{NP}$einen Gewinn] [$_{PP}$von 150 Millionen DM],

[$_{Comp}$weil] [$_{NP}$die Auftraege] [$_{PP}$im Vergleich] [$_{PP}$zum Vorjahr] [$_{Card}$um 13%] [$_{V}$gestiegen sind].

"The siemens company has made a revenue of 150 million marks in 1988, since the orders increased by 13% compared to last year."



Das Referenzresolutionsmodul (RRM) operiert auf der XML-Ausgabe von SMES



Blinde Evaluation der aktuellen IE-Kernkomponenten (Korpus: Pressemitteilungen)

Lexikalische Komponente (20.000 Tokens)

	Recall %	Precision %
Kompositanalyse	99.01	99.29
Part-of-Speech-Filter	95.50	97.90
Namen (inkl. dynam. Lexikon)	85.00	95.77
Fragmente (NPs, PPs):	76.11	91.94

Divide-and-conquer Parser (400 Sätze, 6306 Wörter)

Verb-Modul	98.10	98.43	
Base-Clause-Modul	93.08 (94.61)	93.80 (93.89)	
Main-Clause-Modul	89.00 (93.00)	94.42 (95.62)	
Gesamtanalyse	84.75	89.68	F=87.14

Systemimplementation und Performanz

Lisp Version von divide-and-conquer Parser

0.57 sec/Satz, Satzlänge 26 Wörter (Durchschnitt)

C++ Re-implementation (Lexikonkomponente & Fragmente & Verb-Modul)

1.2 MB von „Wirtschaftswoche“ (197118 Tokens): ~32sec

~6160 wrds/sec; PentiumIII, 500MHz, 128Ram

> Faktor 20 verglichen mit Lisp-Version (~ 75 -100 Sätze/sec)

Aktueller Einsatzbereich am DFKI:

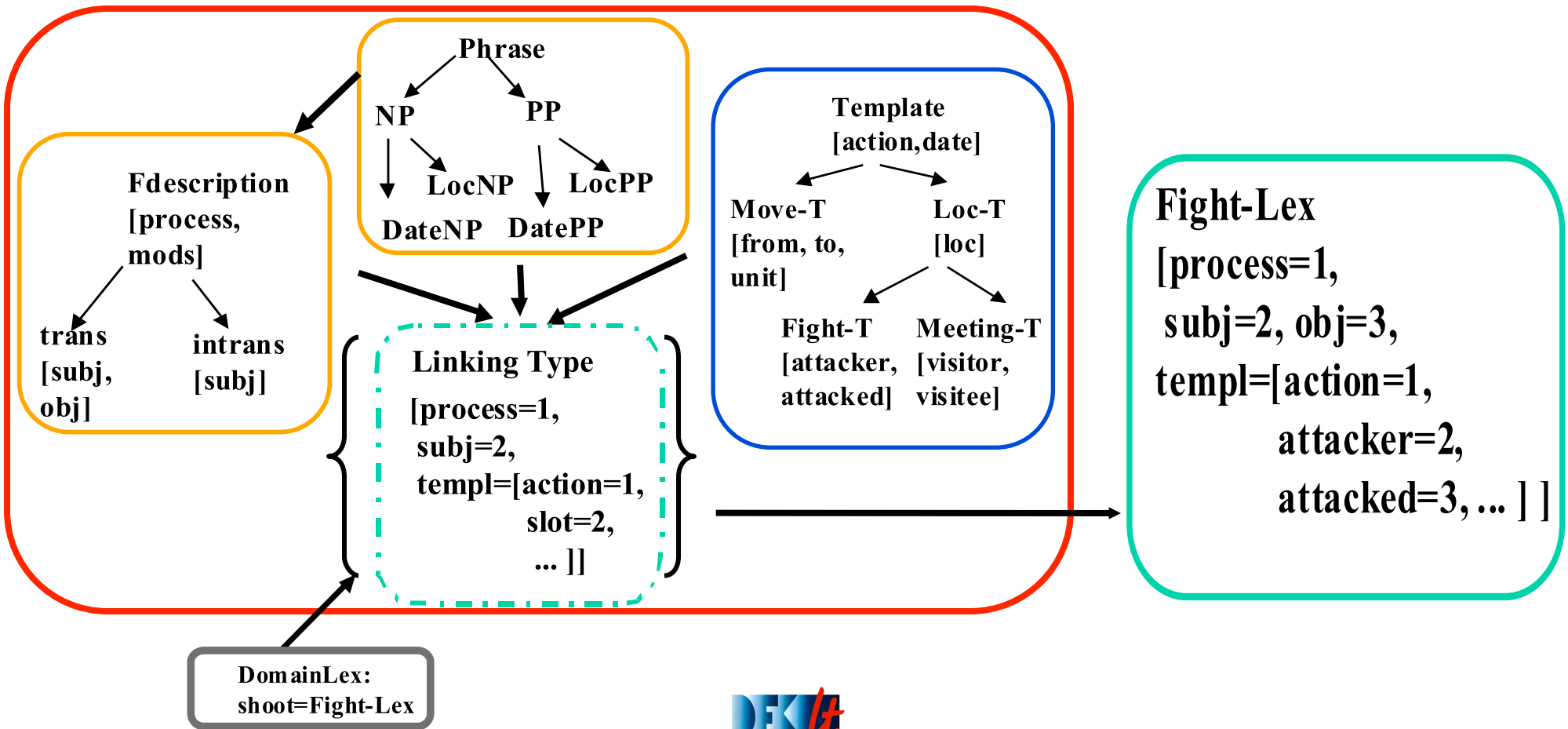
Umsatzmeldungen, Call-Center, Textklassifikation, Ontologie-Extraktion,
NL-Anfrage, Suchmaschinen, Extraktion von Video-Untertiteln,
Textzusammenfassungen

Multi-Lingualität:

Englisch (im Rahmen des Getess Projektes),
erste (kleinere) Demonstratoren für Chinesisch, Japanisch

Domänenmodellierung wird in SMES mittels getypter Merkmalsstrukturen realisiert

- ✓ Die Domänemodellierung ist mittels hierarchisch organisierter Templates realisiert (blaue Box), unter Verwendung von TDL, womit auch Generalisierungen von linguistischen Objekten definiert sind (gelbe Boxen).
- ✓ Das Interface zwischen Domänenwissen und Linguistik wird mittels *linking types* (grüne Box) realisiert, die eine Kombination zwischen Konzepten der verschiedenen Ebene darstellen, und über ein Domänenlexikon (graue Box) zugreifbar sind (grüne Box). Template-Filling wird dann durch Typexpansion geleistet.



Dieses Model hat mehrere Vorteile

- Spezifikation neuer Anwendungen im wesentlichen durch
  - Template-Hierarchie (unabhängig von linguistischen Wissenquellen)
  - Integration von Domänen-und Sprachwissen durch linking types zwischen abstrakten Kategorien

- Unterstützung einer schnellen Adaption an neue Anwendungen

- Weitere Vorteile:
  - Systematische Integration von lexikalischer Semantik (word net)
  - Klare Schnittstellen für automatische Wissenserwerbsmethoden
    - Lernen von Domänenlexikon
    - Lernen von linking types

Aktuell gewinnen Methoden des Maschinellen Lernens zum (semi-) automatischen Erwerbs von IE-Routinen an Bedeutung

- Die Mehrzahl der aktuellen Methoden sind dabei Varianten des überwachten induktiven Lernens

- Ziele: ausgehend von einer Menge domänenspezifisch annotierter Textdokumente werden automatisch Templateinstanziierungsregeln erworben durch sukkzessive Generalisierung der instanziierten, initialen Regelmenge, die mittels der Trainingsmenge bestimmt wurde.

- Dokumente werden mittels NL-Komponenten vorverarbeitet
  - Tokenization (Freitag, 98)
  - POS tagging (Califf&Mooney,98)
  - Phrasenerkennung (Hufman,96)
  - Flache Satzanalyse (Riloff,96a;Soderland,97)

- Die meisten Ansätze lernen Slot-Füller-Regeln, einige neuere Ansätze bereits auch relationale Strukturen (Califf&Mooney,98, Miller et al. 2000, Neumann:in prep)

- Der gegenwärtige Trend geht auch in Richtung minimal überwachter Lernstrategien (Riloff,96b, Yangarber et al. 2000)

Die Mehrzahl der aktuellen Methoden sind Varianten des überwachten induktiven Lernens

- Beispiel

  <PNG> Sue Smith </PNG>, 39, of Menlo Park, was appointed <TNG> president </TNG> of <CNG> Foo Inc. </CNG>

  n_was_named_t_by_c:

  noun-group(PNG, head(isa(person-name))),
  noun-group(TNG, head(isa(title))),
  noun-group(CNG,head(isa(company-name))),
  verb-group(VG, type(passive), head(named or elected or appointed)),

  prep(PREP, head(of or at or by)),
  subject(PNG,VG), object(VG,TNG), post_nominal_prep(TNG,PREP), prep_obj(PREP,CNG)

  ⇒ management_appointment(M,person(PNG), title(TNG), company(CNG))

- Die aktuellen Forschungsergebnisse sind bereits recht vielversprechend (für flache Satz-basierte Templates):
  - Hufman,96: management changes task => 85.2% F (89.4%)
  - Califf&Mooney,98 : computer-related job postings => 87.1% P & 58.8% R

Auch im Falle der multi-lingualen Erkennung von Eigennamen sind maschinelle Lernverfahren sehr vielversprechend

- Nymble™, a high performance learning name-finder (Bikel et al. 97, BBN)
  - Hidden-Markov Model
  - Wortmerkmale (z.B., allCaps, twoDigitNum)
  - Resultate gemäß F-measure
    - Englisch: 93% (bestes MUC-6: 96%)
    - Spanisch: 90% (93%)

- Gallippi, 96:
  - Datengesteuerte Strategie basierend auf Entscheidungsbäumen (ID3)
  - Merkmale: POS, Abkürz., Namenslisten, Wortmerkmale
  - Resultate gemäß F-measure (Mittel über Firmen-, Personen-, Ortsnamen)
    - Englisch: 94%
    - Spanisch: 89.2%
    - Japanisch: 83.1%

Konstruktion von IE-Anwendungen automatisch auf der Basis von annotierten Textbeispielen und der durch SMES bestimmten XML-Strukturen

- Repräsentation
  - getaggte Templateinstanzen aus Paragraphen als Templates-Trees
  - XML-Ausdrücke von SMES als XML-Trees

- Merging korrespondierender Template-Trees und XML-trees liefert expandierte Template-Trees

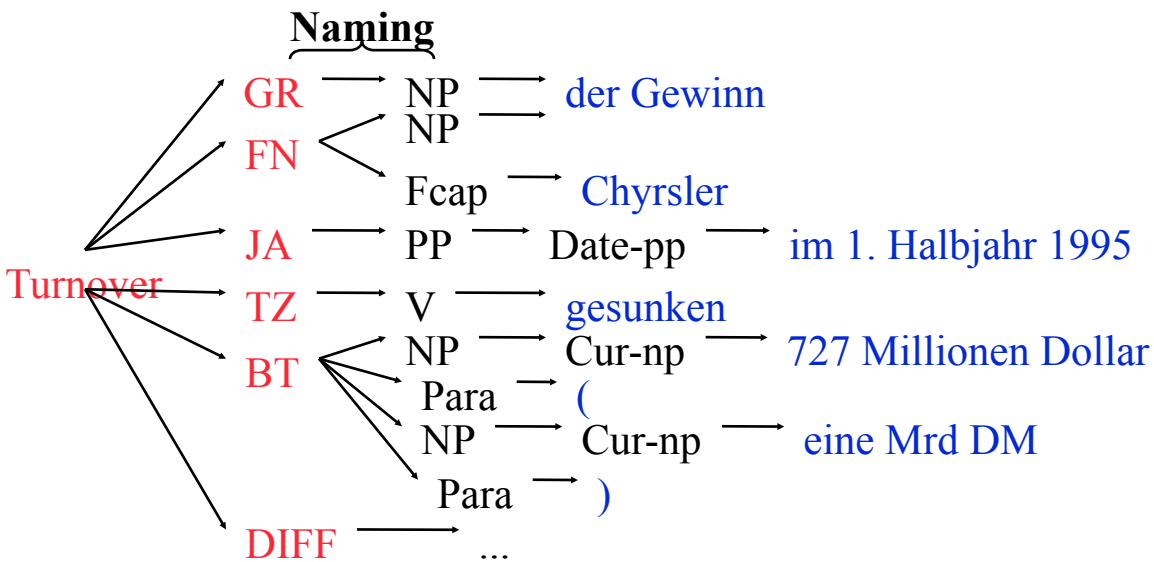
- Dekomposition der expandierten Template-Trees in stochastisch lexikalisierte Tree-Grammatiken (IE-Grammatik)

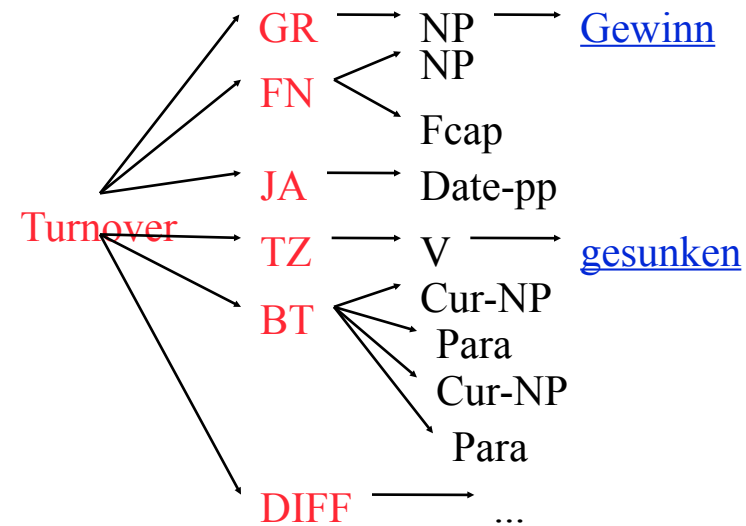
- Informationsextraktion durch Baumkomposition auf Basis der XML-Analyse und IE-Grammatik

- Genetische Programmierung zur Erzeugung neuer Template-Trees

Die Struktur der Templates wird durch getaggte Paragraphen definiert (Paragraphen-orientierte IE)

<GR 1>Der Gewinn</GR> <FN 2><FN 1>des amerikanischen Autoherstellers Chrysler</FN></FN> ist <JA 1>im 1. Halbjahr 1995</JA> deutlich <TZ 1> gesunken </TZ>. Der drittgrößte US-Automobilkonzern hat im Berichtszeitraum nur noch <BT 1>727 Millionen Dollar (eine Mrd DM)</BT> verdient <DIFF 1> gegenüber 1,9 Milliarden Dollar (2,65 Mrd DM) im 1. Halbjahr 1994 </DIFF>. <GR 2>Der Umsatz</GR> <TZ 2>ging</TZ> geringfügig <DIFF 2>um 200 Millionen Dollar</DIFF> <BT 2>auf 26,1 Milliarden Dollar</BT> <TZ 2>zurück</TZ>.





Zusammenfassend: IE ist ein sehr attraktives Forschungsgebiet zur Konstruktion von Anwendungssystemen

- IE ist interdisziplinär
  - Sprachtechnologie
  - Statistische Methoden
  - Maschinelles Lernen
  - Wissensrepräsentation
  - Software-Engeenering
  - Expertenwissen

- Kommende Systemgenerationen
  - werden noch intelligenter sein durch Selbstadaption an neue Domänen
  - aus Erfahrung lernen mittels minimal überwachter Methoden/User Feedback
  - multi-mediale Quellen zur Extraktion und Präsentation
  - IE-basierte multi-agenten Suchmaschinen/Konzeptagenten
  - IE-basierte Frage/Antwortsysteme (Speech-Interfaces)